DOCUMENT RESUME

ED 189 161 TM 800 332

AUTHOR Reynolds, Cecil R.; And Others
TITLE A Regression Analysis of Test Bias on the Stanford-Binet Intelligence Scale.
PUB DATE Apr 80
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Black Students; Elementary Education; *Intelligence Tests; Learning Problems; *Predictive Validity; *Racial Differences; *Regression (Statistics); *Test Bias; White Students
IDENTIFIERS *Stanford Binet Intelligence Scale; Wide Range Achievement Test

ABSTRACT

Regression lines for the prediction of Wide Range Achievement Test (WRAT) standard scores by Stanford-Binet Intelligence Scale scores were compared across race for matched groups of 60 black and 60 white children selected from among a large number of children who had been referred for psychological services by their classroom teachers. The white children were matched to the black children on the basis of sex, age, and IQ. The definition proposed for bias was significant differences in regression lines using Potthoff's technique that tests for slopes and intercepts simultaneously. According to this significance test, regression lines for blacks and for whites did not differ significantly for the prediction of WRAT scores. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189161

A Regression Analysis of Test Bias on the Stanford-Binet Intelligence Scale

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Cecil R. Reynolds
Acting Director
Buros Institute of Mental Measurements
135 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

Michael D. Bossard
Doctoral Candidate
School Psychology Training Program
130 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

Terry B. Gutkin
Director
School Psychology Training Program
130 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Reynolds
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A paper presented to the annual meeting of the American Educational Research Association, Boston: April, 1980.

TM 800332

Abstract

Regression lines for the prediction of WRAT standard scores by Stanford-Binet IQs were compared across race, by the Potthoff procedure, for equated groups (sex, age, and IQ) of 60 Black and 60 White children referred for psychological services by their classroom teachers. Regression lines for Blacks and Whites did not differ significantly for the prediction of WRAT scores by the Stanford-Binet. Implications of these findings are discussed.

The use of psychological tests normed primarily with white children for psychological diagnosis and educational decision-making concerning minority children, has become a very hotly contested issue in recent years. Although much discussion of the issue has appeared in both the scientific and public literature, few data of relevance to the issue (with school age children) have been presented. The use of such tests are of special concern to psychologists involved in assessment, particularly in view of the Larry P. case (Note 1) and P. L. 94-142 (Note 2; Education for All Handicapped Children Act of 1975). Harrington (1976, 1975) has gone so far as to state that it is not possible for tests developed and normed on a white majority to be other than biased against minorities and to show less predictive validity when used with minorities.

In response to pressure from the Black Psychological Association (which was actually requesting a moratorium on the use of psychological and educational tests with disadvantaged students), the APA Board of Directors requested, in 1968, the Board of Scientific Affairs to appoint a group to study the use of such tests with disadvantaged students. In reporting on this issue, the committee (Cleary, Humphreys, Kendrick, & Wesman, 1975) offered a definition of test bias. While including content and construct validity as important variables in the issue of test bias, the focus was clearly

on predictive validity:

> A test is considered fair for a particular use if the inference drawn from the test score is made with the smallest feasible random error and if there is no constant error in the inference as a function of membership in a particular group. (Cleary et al., 1975, p.25)

The definition of bias offered by the APA commmittee is a restatement of previous definitions by Cardall and Coffman (1964), Cleary (1968), and Potthoff (1966), and has been widely accepted (though certainly not without criticism, e.g., Bernal, 1975; Linn & Werts, 1971; Thorndike, 1971). Oakland and Matuszek (1977) examined class placement procedures under several proposed models of bias and demonstrated that the Cleary model results in the smallest number of children being misplaced, although under certain legislative conditions, Oakland and Matuszek favored the Thorndike (1971) "quota" selection model. After reviewing a number of models, Peterson and Novick (1976) designated the regression model as the most logically tenable and the most widely used placement model. A statistical technique provided by Potthoff (1966) has also received widespread acceptance in the examination of regression lines to test bias under the Cleary et al. definition (Schmidt & Hunter, 1974).

While considerable data are available on the validity of the Scholastic Achievement Test (e.g., Goldman & Hewitt, 1976; Kallingal, 1971; Pfiefer & Sedlacek, 1971) and various employment tests (e.g. Boehm, 1972; Hunter, Schmidt, & Hunter, 1979) for

blacks and whites, only recently have studies appeared dealing with differential validity of IQ tests. Mitchell (1967) studied the validity of two broad based readiness tests to predict first grade achievement for blacks and whites finding similar validity coefficients for the two races. Mitchell's study was limited to comparing the magnitude of independent-dependent variable correlation and did not look for identity of regression lines. Hartlage, Lucas, and Godwin (1976) compared the predictive validity of the WISC and Raven with a group of low SES, disadvantaged children. When comparing what they considered to be the relatively culture-fair test, the Raven Matrices, with the "culture-loaded" 1949 WISC, Hartlage et al. (1976) found the WISC to have consistently larger correlations with measures of reading, spelling, and arithmetic than the Raven. These authors only compared the strength of the relationship in each case and did not look for identity of regression lines (equivalent beta coefficients and intercept constants).

More recently, Reynolds and Hartlage (1979) compared regression lines for the prediction of achievement by the WISC and the WISC-R across race for blacks and whites. Their results indicated that regression lines for blacks and whites did not differ significantly. Reynolds and Gutkin (1980) replicated the Reynolds and Hartlage (1979) study for the WISC-R, comparing regression lines between whites and Mexican-Americans. Again, no significant differences

were found. In a study with much larger samples, Reschly and Sabers (1979) investigated the WISC-R's ability to predict Metropolitan Achievement Test scores across four ethnic groups (blacks, whites, chicanos, and native American Papagos). Reschly and Sabers (1979) adopted the Cleary, regression definition and a procedure by Gulliksen and Wilks (1950) that separately tests slopes and intercepts (whereas the Potthoff, 1966, technique simultaneously tests slopes and intercepts). They found that the WISC-R was for the most part equally valid for the different groups. When differences occurred, they were due to variations in intercepts resulting in the over-prediction of performance for non-white groups.

The purpose of the present study is to provide data that will aid in the empirical evaluation of test bias (under the Cleary et at., 1975, definition) for the Stanford-Binet Intelligence Scale, Form L-M, 1972 Norms Edition (Terman & Merrill, 1973). It was hypothesized that, as with previous research on the WISC and WISC-R, no significant differences would occur between regression lines across groups. Previous research on bias has ignored the Binet. The Binet should be of particular interest in test bias research since it has historically been the IQ test against which new tests have been validated.

METHOD

Subjects

The sample consisted of equated groups of 60 white and 60 black urban children referred by teachers for psychological evaluation due to a variety of learning and/or behavior problems. A referral population was chosen because they are the predominant group of interest in the prediction of achievement from the IQ. The children were chosen as follows from more than 1,000 district-wide referrals. A computer listing of all children with complete data was obtained. Every third black male was chosen until 30 children were obtained. The procedure was repeated for black females. Since random assignment to race or sex is not possible, whites were chosen to match the black children on the variables of age (within 6 months), sex, and IQ (within 10 points). To match the groups, a black child was chosen and records of the white group examined. The first matching white child to be encountered was selected. The resulting sample characteristics are described in greater detail in Table 1. The relatively low IQ of the groups is typical of referral populations (Gutkin & Reynolds, in press; Reynolds, Gutkin, Dappen, & Wright, 1979; Reynolds & Hartlage, 1979).

Insert Table 1 about here

Procedure

The Stanford-Binet Intelligence Scale (Terman & Merrill, 1973) and the most recent revision of the Wide Range Achievement Test (Jastak & Jastak, 1978) were administered by certified school psychologists and psychological assistants. Testing on both scales was accomplished during a single session.

Regression lines for each pair of scores (Binet IQ predicting each WRAT subtest) were examined across race through the Potthoff (1966) technique. This procedure yields a single F - ratio that simultaneously tests regression coefficients (slopes) and intercept values. If a significant F results, slopes and intercepts may then be assessed separately to determine whether the resulting bias in prediction is constant (intercepts differ) or changes with the distance of scores from the mean (slopes differ). Slopes and intercepts must both be equivalent prior to concluding homogeneity of regression across groups. Only when slope and intercepts are the same can a common regression equation (derived by combining the groups in question) be applied. If homogeneity of regression across groups does not occur, then in order to have fair use of test scores, separate equations for each group must be employed.

RESULTS

Regression lines for blacks and whites did not differ at the .05 level of significance for the prediction of WRAT Reading, F (2,116) = 1.24 $p > .05$, Spelling, F (2,116) = 0.18, $p > .05$, or

Arithmetic, F (2,116) = 2.24 $p > .05$, standard scores by the Stanford-Binet IQ. Thus, present results provide support for the use of a common regression equation (Bossard & Galusha, in press) to predict WRAT achievement scores for referred black and white children with the Stanford-Binet. Correlations between the Stanford-Binet IQ and achievement for both groups were quite substantial, never accounting for less than 49% of the variance in achievement scores. For black children the correlations were: .74 with Reading, .78 with Spelling, and .70 with Arithmetic. For whites the correlations were: .81 with Reading, .81 with Spelling, and .82 with Arithmetic. As expected from the results of the Potthoff analysis, the pairs of correlations are quite similar across these two racial groupings.

DISCUSSION

The study's results are consistent with previous investigations of test bias using the regression definition. That is, standardized intelligence tests have been shown to predict school achievement about equally well for blacks and whites. Prior to concluding that the Stanford-Binet Intelligence Scale is free of bias in terms of predictive accuracy (the regression definition), more research is needed utilizing a wide variety of criterion measures including other individual achievement tests, group achievement tests, and teacher constructed scales. Studies of this kind will help to evaluate the relative influence of bias within different criterion

measures. Since using a referral population may minimize differences between groups, replication with normal children will also need to be undertaken.

Test developers need to become more aware of the issue of bias, to the point of demonstrating validity across groups prior to publication of the instrument. While this has occurred somewhat in the area of achievement testing (Anastasi, 1976), investigations of differential validity by test publishers are conspicuously lacking. Studies similar to the present investigation are needed with other existing measurement instruments to determine whether alterations in interpretation of the scales are needed when applied to groups other than the majority population.

At present however, a considerable body of data is accumulating indicating consistency of content (Jensen & Figueroa, 1975), construct (Gutkin & Reynolds, in press; Jensen, 1976; Reschly, 1978; Reynolds, in press a,b), and predictive (Reschly & Sabers, 1979; Reynolds & Gutkin, 1980; Reynolds & Hartlage, 1979) validity of the IQ test across racial groupings.

Reference Notes

1. Larry P. *et al*. vs. Wilson Riles *et al*., 343 F. Supp. 1306 (D.C.N.D. Cal., June 20, 1972).

2. The Education for All Handicapped Children Act of 1975, Pub. L. No. 94-142, 89 stat. 773.

References

Anastasi, A. Psychological Testing. 4th Ed. New York: MacMillan, 1976.

Bernal, E. M. A response to "Educational uses of tests with disadvantaged students." American Psychologist 1975, 30, 93-95.

Boehm, V. Negro-white differences in validity of employment and training selection procedures: Summary of recent evidence. Journal of Applied Psychology, 1972, 56, 33-39.

Bossard, M. D., & Galusha, R. The utility of the Stanford-Binet in predicting WRAT performance. Psychology in the Schools, in press.

Cardall, C. & Coffman, W. A method for comparing the performance of different groups on the items in a test. Research and Development Reports, 1964, 64-5, No. 9, College Entrance Examination Board.

Cleary, T. A. Test Bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 15-41.

Goldman, R. & Hewitt, B. Predicting the success of black, chicano, oriental, and white college students. Journal of Educational

Measurement, 1976, 13, 107-117.

Gulliksen, J., & Wilks, S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Gutkin, T. B., & Reynolds, C. R. Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. Journal of School Psychology, in press.

Harrington, G. M. Minority test bias as a psychometric artifact: The experimental evidence. Paper presented at the symposium, Race and sex differences in ability, at the annual meeting of the American Psychological Association, Washington: September, 1976.

Harrington, G. M. Intelligence tests may favour the majority groups in a population. Nature, 1975, 258, 708-709.

Hartlage, L. C., Lucas, T., & Godwin, A. Culturally biased and culture fair tests correlated with school performance in culturally disadvantaged children. Journal of Clinical Psychology, 1976, 32, 235-237.

Hunter, J. E., Schmidt, F. L. & Hunter, R. Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 1979, 86, 721-735.

Jastak, J. F., & Jastak, S. R. Manual. The Wide Range Achievement Test. (Rev. Ed.). Wilmington, DE: Guidance Associates of Delware, Inc., 1978.

Jensen, A. R. Test bias and construct validity. Phi Delta Kappan, 1976, 58, 340-346.

Jensen, A. R., & Figueroa, R. A. Forward and backward digit-span interaction with race and IQ. Journal of Educational Psychology, 1975, 67, 882-893.

Kallingal, A. The prediction of grades for black and white students at Michigan State University. Journal of Educational Measurement, 1971, 8, 263-265.

Linn, R. L. & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

Mitchell, B. C. Predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrel Reading Readiness Analysis for negro pupils. Educational and Psychological Measurement, 1967, 27, 1047-1054.

Oakland, T., & Matuszek, P. Using tests in non-discriminatory assessment. In T. Oakland (Ed.), Psychological and educational assessment of minority group children, NYC: Brunner/Mazel, 1977.

Peterson, N. & Novick, M. An evaluation of some models for culture fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Pfiefer, C. & Sedlacek, W. The validity of academic predictors for black and white students at a predominantly white university. Journal of Educational Measurement, 1971, 8, 253-261.

Potthoff, R. F. Statistical aspects of the problem of bias in psychological tests. Institute of Statistics Mimeo Series No. 479, Chapel Hill, N. C.: UNC-Chapel Hill Department of Statistics, 1966.

Reschly, D. Nonbiased assessment. In G. Phye & D. Reschly (Eds.) School Psychology: Perspective and Issues. New York: Academic Press, 1979.

Reschly, D., & Sabers, D. Analysis of test bias in four groups with the regression definition. Journal of Educational Measurement, 1979, 16, 1-9.

Reynolds, C. R. Differential construct validity of a preschool battery for blacks, whites, males, and females. Journal of School Psychology, in press. a.

Reynolds, C. R. The invariance of the factorial validity of the Metropolitan Readiness Tests. Educational and Psychological Measurement, 1979, in press. b.

Reynolds, C. R., & Gutkin, T. B. A regression analysis of test bias on the WISC-R for Anglos and Chicanos referred for psychological services. Journal of Abnormal Child Psychology, 1980, in press.

Reynolds, C. R., Gutkin, T. B., Dappen, L., & Wright, D. Differential validity of the WISC-R for boys and girls referred for psychological services. Perceptual and Motor Skills, 1979, 48, 868-870.

Reynolds, C. R., & Hartlage, L. C. Comparison of WISC and WISC-R regression lines for academic prediction with black and with white referred children. Journal of Consulting and Clinical Psychology, 1979, 47, 589-591.

Schmidt, F., Berner, J., & Hunter, J. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 58, 5-9.

Terman, L. M. & Merrill, M. A. Stanford-Binet Intelligence Scale. Boston: Houghton Mifflin, 1973.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 7, 63-70.

Table 1

Sample Characteristics by Race and Sex

			Age in Years		Stanford-Binet IQ	
	Sex	N	$\bar{X}$	SD	$\bar{X}$	SD
Blacks	M	30	8.38	2.56	82.82	17.23
	F	30	8.53	2.70	83.33	20.79
Whites	M	30	8.30	2.79	84.53	16.68
	F	30	8.42	2.88	84.16	23.99

Wide Range Achievement Test

			Reading		Spelling		Arithmetic	
	Sex	N	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Blacks	M	30	83.43	16.11	83.83	18.16	84.07	16.42
	F	30	83.30	16.47	84.20	17.37	82.50	17.95
Whites	M	30	82.97	16.63	84.77	19.62	80.83	19.07
	F	30	84.90	23.48	85.77	20.66	83.47	23.22